John von Neumann Institute for Computing

Modeling Conformational and Molecular Weight Heterogeneity with Analytical Ultracentrifugation Experiments

B. Demeler, E. Brookes

published in

From Computational Biophysics to Systems Biology (CBSB08),
Proceedings of the NIC Workshop 2008,
Ulrich H. E. Hansmann, Jan H. Meinke, Sandipan Mohanty,
Walter Nadler, Olav Zimmermann (Editors),
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **40**, ISBN 978-3-9810843-6-8, pp. 73-76, 2008.

http://www.fz-juelich.de/nic-series/volume40

Modeling Conformational and Molecular Weight Heterogeneity with Analytical Ultracentrifugation Experiments

Borries Demeler[1] **and Emre Brookes**[2]

[1] The University of Texas Health Science Center at San Antonio,
Department of Biochemistry, San Antonio, Texas, USA
E-mail: demeler@biochem.uthscsa.edu

[2] The University of Texas at San Antonio,
Department of Computer Science, San Antonio, Texas, USA
E-mail: ebrookes@cs.utsa.edu

Sedimentation velocity experiments reveal information about molecular weight and shape of sedimenting macromolecules. The observables in such experiments are the sedimentation and diffusion coefficients and the concentration of individual solutes. We have developed parallel optimization algorithms that allow us to extract molecular parameters from mixtures of macromolecules using a nearly model-independent approach. Using a combination of deterministic and stochastic optimization, we are able to fit complex analytical ultracentrifugation experiments globally with excellent convergence properties. Our software uses the TIGRE grid middleware to distribute the computing effort to Teragrid and other computing resources, and offers a public web portal for the hydrodynamic analysis of AUC experiments[1]. Our solutions provide unparalleled resolution, and allow us to characterize polymerization events, aggregation and provide high resolution information in structure and function studies in the solution state.

1 Introduction

The sedimentation and diffusion transport of a solute observed in an analytical ultracentrifugation (AUC) sedimentation velocity experiment is described by the Lamm equation[2]. Mixtures of solutes can be modeled well by a linear combination of finite element solutions of the Lamm equation[3,4] where each term represents a different solute in the mixture. The sedimentation (s_k) and diffusion coefficients (D_k) are parameters of the Lamm equation, and define uniquely the molecular weight and shape of each solute k in the mixture, while the amplitude of each term determines the partial concentration (c_k). In an AUC experiment the goal is to correctly determine s, D, c as well as n, the number of solutes present in the mixture. The inverse problem of fitting experimental data to simulations of Lamm equation systems represents a difficult optimization problem which is nonlinear with respect to the fitting parameters. We present here a method for evaluating experimental data by applying multiple optimization algorithms in series for obtaining the most likely parsimonious parameter distribution that satisfies Occam's razor. Our approach is implemented on a parallel computing platform utilizing the globus-based TIGRE grid middleware[5] which can be conveniently accessed through a web portal. Results can be further analyzed with the UltraScan software[6,7]. Our approach includes algorithms for initialization, systematic noise deconvolution, parameter search and parsimonious regularization.

2 Initialization

The parameter search requires an initialization step which identifies the limits of the domains of two of the fitting parameters, s, and D. If the experimental data contain significant amount of time invariant systematic noise, the s limits are conveniently identified with the time-derivative method by Stafford, which yields a model-free transformation of the primary data that eliminates any time invariant noise contributions[8]. A more accurate initialization can be obtained from the experimental data by the enhanced van Holde - Weischet method[9], which yields a model-independent, diffusion-corrected sedimentation distribution for cases where time invariant noise is not significant. Accurate limits for D are difficult to obtain reliably by model-independent means, and require prior knowledge and parameterization by the frictional ratio, f_r:

$$D_k = \frac{RT}{18\pi N}\left[\frac{2(1-\bar{\nu}_k\rho)}{s\,\bar{\nu}}\right]^{\frac{1}{2}}(\eta f_{r,k})^{-\frac{2}{3}} \tag{1}$$

where R is the gas constant, T the temperature in Kelvin, N is Avogadro's number, η and ρ are the viscosity and density of the solvent, and $\bar{\nu}_k$ is the partial specific volume of solute k. Values for f_r are chosen based on the analyte under investigation, for example 1-2 for globular macromolecules, 1.5-3 for disordered or denatured proteins, or values up to 10 for elongated molecules such as long nucleic acids, fibrils or amyloid aggregates. For unknown systems a sufficiently large range can also be chosen, but in those cases additional refinement steps may be required.

3 Time-Invariant Noise Reduction and 2-Dimensional Spectrum Analysis Parameter Search

The precision of parameter estimation is inversely correlated with the experimental noise present in the primary data. It is therefore important that systematic noise contributions resulting from instrument flaws are accounted for and that stochastic noise contributions are attenuated using Monte Carlo (MC) methods[10]. We have shown that systematic noise contributions can be effectively eliminated using algebraic means[11]. Experimental design considerations can further improve noise characteristics, for example, by using intensity measurements instead of absorbance measurements, stochastic noise is reduced by a factor of $\approx \sqrt{2}$ by not subtracting the reference signal. This subtraction leads to the convolution of two stochastic noise vectors and an increase in the stochastic noise. In the first optimization step we perform a 2-dimensional spectrum analysis between the limits determined above in section 2 as described by Brookes et al.[12]. Briefly, a divide and conquer algorithm is employed to search multiple coarse-grained subgrids spanning the entire 2-dimensional parameter range in s and f_r. Each grid point is an element in a linear combination of finite element solutions of the Lamm equation, whose amplitudes are determined in a least squares fit using the non-negatively constrained linear least squares (NNLS) fitting algorithm[13]. By combining the results from multiple relatively coarse grids that are slightly offset against each other, a high-resolution, 2-dimensional spectrum analysis (2DSA) is obtained. The result is a sparse matrix identifying potential signals in the sample.

4 Parsimonious Regularization of the 2DSA Grid Using Genetic Algorithms

After performing the 2DSA analysis, a sparse grid identifying potential solutes is obtained. However, due to the presence of experimental noise and due to the degeneracy resulting from fitting with an overdetermined system the result is subject to the presence of false positives. While such effects cannot be entirely eliminated, a parsimonious regularization can improve the solution significantly by providing a solution that satisfies Occam's razor. Occam's razor states that the solution with the greatest parsimony of parameters resulting in nearly the same residual mean square deviation (RMSD) as another less parsimonious solution is to be preferred. We have implemented a second step in the optimization process which takes advantage of a genetic algorithm (GA) approach. In this approach, we initialize the GA analysis with parameter constraints obtained by drawing 2-dimensional boundaries with user-defined width around each solute. Overlaps between adjacent boxes are eliminated by further subdividing existing boxes to create new, non-overlapping boxes. During fitting, parameters are adjusted in an evolutionary approach based on fitness:

$$fitness = RMSD * \left(1 + (rf * nz(x))^2\right) \quad (2)$$

where nz is the cardinality of solution x and rf is a regularization factor applying RMSD penalties to increase parsimony[14].

5 Global Multi-Speed Genetic Algorithm Monte Carlo Refinement

In order to enhance the information content of AUC experiments, data from multiple experiments performed at different speeds can be combined in a global fit. In high speed experiments, sedimentation signals are enhanced, in low speed experiments diffusion signals are improved. GA-MC analysis can be performed globally by constraining the solute model to all datasets. Table 1 lists results from a simulated 5-component system with heterogeneity in both shape and molecular weight (realistic noise added), representing a linear elongation event, performed at both 20 krpm and 60 krpm.

Solute	Molecular Weight (kD)	Partial Concentration	Frictional Ratio, f/f0
1	24.26 (24.20, 24.33) [25]	0.0972 (0.0966, 0.0982) [0.1]	1.21 (1.21, 1.21) [1.2]
2	48.04 (47.74, 48.46) [50]	0.102 (0.101, 0.104) [0.1]	1.41 (1.40, 1.42) [1.4]
3	100.2 (97.96, 101.8) [100]	0.0995 (0.0982, 0.101) [0.1]	1.65 (1.63, 1.67) [1.6]
4	198.0 (194.2, 200.8) [200]	0.0996 (0.0989, 0.101) [0.1]	1.84 (1.82, 1.86) [1.8]
5	385.3 (380.4, 394.0) [400]	0.100 (0.100, 0.101) [0.1]	2.01 (1.99, 2.04) [2.0]

Table 1. GA-MC results for a global fit of a multispeed 20/60 krpm experiment (described in text). The results demonstrate remarkable agreement with the target. Parentheses: 95% confidence intervals; square brackets: target value. All values rounded off to 3 or 4 significant digits.

Acknowledgments

We would like to thank the NIH-NCRR for funding this research through Grant RR022200, and the NSF for Teragrid allocation TGMCB070038 (both to B.D.).

References

1. Demeler, B. 2006. *A web-based Laboratory Information Management System for UltraScan*. `http://www.uslims.uthscsa.edu`.
2. Lamm, O., 1929. *Die Differentialgleichung der Ultrazentrifugierung*. Ark. Mat. Astron. Fys. **21B**, 1–4, 1928.
3. Cao, W. and B. Demeler. *Modeling analytical ultracentrifugation experiments with an adaptive space-time finite element solution of the Lamm equation*. Biophys. J. **89**, 1589–1602, 2005.
4. Cao, W. and B. Demeler. *Modeling Analytical Ultracentrifugation Experiments with an Adaptive Space-Time Finite Element Solution for Multi-Component Reacting Systems*. Biophys. J. **95(1)**, 54–65, 2008.
5. The Texas Internet Grid for Research and Education. A Globus-based grid middleware for sharing computational resources across a wide area network, `http://www.tigre.net`.
6. Demeler, B. (2008) UltraScan: A Comprehensive Data Analysis Software Package for Analytical Ultracentrifugation Experiments. Version 9.7, `http://www.ultrascan.uthscsa.edu`.
7. Demeler, B. (2005) *UltraScan A Comprehensive Data Analysis Software Package for Analytical Ultracentrifugation Experiments*, in Modern Analytical Ultracentrifugation: Techniques and Methods (D. J. Scott, S. E. Harding, and A. J. Rowe, Eds.) 210-229, Royal Society of Chemistry (UK).
8. Stafford, W. F. *Boundary analysis in sedimentation transport experiments: a procedure for obtaining sedimentation coefficient distributions using the time derivative of the concentration profile*. Anal. Biochem. **203**, 1–7, 1992.
9. Demeler, B. and K. E. van Holde *Sedimentation velocity analysis of highly heterogeneous systems*. Anal Biochem **335(2)**, 279–88, 2004.
10. Demeler, B. and E. Brookes. *Monte Carlo analysis of sedimentation experiments*. Colloid Polym Sci **286(2)**, 129-137, 2008.
11. Schuck P and Demeler B. *Direct sedimentation analysis of interference optical data in analytical ultracentrifugation*. Biophys J. **76(4)**, 2288-96, 1999.
12. Brookes, E. H., R. V. Boppana and B. Demeler. *Computing Large Sparse Multivariate Optimization Problems with an Application in Biophysics.* SuperComputing 2006 Conference Proceedings 0-7695-2700-0/06 2006 IEEE (`http://sc06.supercomputing.org/schedule/pdf/pap320.pdf`).
13. Lawson, C. L. and Hanson, R. J. 1974. *Solving Least Squares Problems.* Prentice-Hall, Inc. Englewood Cliffs, New Jersey.
14. Brookes, E and B. Demeler. *Parsimonious Regularization using Genetic Algorithms Applied to the Analysis of Analytical Ultracentrifugation Experiments*. GECCO Proceedings ACM 978-1-59593-697-4/07/0007 (2007).